DeepAmp: A Convolutional Neural Network Based Tool for Predicting Protein AMPylation Sites From Binary Profile Representation

Sayed Mehedi Azim
United International University
Alok Sharma
Griffith University
Swakkhar Shatabda
United International University
Abdollah Dehzangi (i.dehzangi@rutgers.edu)
Rutgers University

Research Article

Keywords: DeepAmp, Neural Network, AMPylation, binary profile

Posted Date: October 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1013130/v1

DeepAmp: A Convolutional Neural Network based tool for predicting protein AMPylation sites from binary profile representation

Sayed Mehedi Azim[1], Alok Sharma[2,3], Swakkhar Shatabda[1,*], and Abdollah Dehzangi[4,5,*]

[1]Department of Computer Science and Engineering, United International University, Plot-2, United City, Madani Avenue, Badda, Dhaka, 1212, Bangladesh
[2]Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia
[3]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan
[4]Department of Computer Science, Rutgers University, Camden, NJ, 08102, USA
[5]Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
[*]corresponding.author@email.example

ABSTRACT

AMPylation is an emerging post-translational modification that occurs on the hydroxyl group of threonine, serine, or tyrosine via a phosphodiester bond. AMPylators catalyze this process as covalent attachment of adenosine monophosphate to the amino acid side chain of a peptide. Recent studies have shown that this post-translational modification is directly responsible for regulation of neurodevelopment and neurodegeneration and also involved in many physiological processes. Despite the importance of this post-translational modification, there is no peptide sequence dataset available for conducting computational analysis. Therefore, so far, no computational approach has been proposed for predicting AMPylation. In this study, we introduce a new dataset of this distinct post-translational modification and develop a new machine learning tool using a deep convolutional neural network called DeepAmp to predict AMPylation sites in proteins. DeepAmp achieves 77.7%, 79.1%, 76.8%, and 0.55 in terms of Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) for AMPylation site prediction task, respectively. As the first machine learning model, DeepAmp demonstrate promising results which highlight its potential to solve this problem. Our presented dataset and DeepAmp as a standalone predictor are publicly available at https://github.com/MehediAzim/DeepAmp

Introduction

Post Translational Modification (PTM) is the enzymic or chemical modification of a protein after it is translated or synthesized in the ribosome. The PTMs are occurred via removal of parts of a translated protein, covalent modifications, or degradation of modified proteins[1,2]. These modifications provide important insight into various cellular functions and biological processes of proteins such as cellular dynamics and elasticity.

PTMs are important mechanisms to increase proteomic diversity, and play a vital role in functional proteomic because they regulate activity, localization, and interaction with other cellular molecules such as proteins, nucleic acids, lipids, and cofactors[3]. They can impact the structure, electrophilicity, and interactions of proteins. PTMs also regulate protein folding via targeting specific subcellular compartments, interacting with ligands or other proteins, or by initiating a change in their functional state including signaling or catalytic activity[4]. A wide range of PTMs have been identified so far. The common PTMs include phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, lipidation, and proteolysis which influence almost all aspects of normal cell biology and pathogenesis[5].

AMPylation is an emerging Post Translational Modification mediated by a bacterial virulence factor that transfers Adenosine Monophosphate (AMP) from Adenosine Triphosphate (ATP) to a threonine residue of eukaryotic substrates[6,7]. AMPylation is the covalent attachment of AMP to a protein or peptide[8]. It has been studied exclusively with the Fic domain proteins, which are preserved and found in proteins stretching from bacteria to humans. By adding AMP to Rho-family GTPases, these enzymes can thereby mediate both bacterial pathogenesis and eukaryotic signaling[9,10]. The most common and stable form of AMPylation occurs on the hydroxyl group of threonine, serine, or tyrosine via a phosphodiester bond. In the AMPylation process, Adenosine Monophosphate (AMP) gets covalently attached to the amino acid side chain of a protein molecule. AMPylation involves a phosphodiester bond between a hydroxyl group of the molecule undergoing AMPylation and the phosphate group of the

adenosine monophosphate nucleotide (i.e. adenylic acid) [14]. The enzymes that are capable of catalyzing this process are called AMPylators. Threonine (T) and Tyrosine (Y) amino acids are usual targets of AMPylation while this PTM can sometimes be observed in Serine (S) as well.

Recent proteomics studies demonstrated that this PTM is more omnipresent than generally acknowledged and it is emerging as a significant regulatory mechanism for both eukaryotic and prokaryotic cells. It is impelled in a vast area of biological processes stretching from regulation of nitrogen metabolism in bacteria and regulation of signaling pathways to pathogenesis in several animal species[11–14]. AMPylation has also found to play a significant role in the regulation of neurodevelopment and neurodegeneration[15].

Experimental approaches used to determine PTM sites are expensive, laborious, and time taking. Hence, many studies have been proposed to predict PTM sites using fast and cost effective computational approaches[16–28].

However, to the best of our knowledge, so far no computational approach has been proposed for predicting AMPylation sites of Fic domain protein. One of the main reasons is that there is no AMPylation dataset available to be used for this task. In this study, we are presenting a new dataset of protein AMPylation sites. Furthermore, we also propose a new deep Convolutional Neural Network (CNN) model called DeepAmp for predicting protein AMPylation sites on the newly found dataset of AMP modified proteins. DeepAmp achieves 77.7%, 79.1%, 76.8%, and 0.55 in terms of Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) for AMPylation site prediction task, respectively. As the first machine learning model, DeepAmp demonstrate promising results which highlight its potential to solve this problem. We believe this study will help researchers immensely in terms of mitigating the current research gap in this subject. Our presented dataset and DeepAmp as an standalone predictor are publicly available at https://github.com/MehediAzim/DeepAmp.

Results and Discussion

Evaluation Metrics

In order to ensure standardized evaluation of our model and to provide more insights into our results, we calculate the Accuracy, Sensitivity, Specificity, and Mathews correlation coefficient (MCC) as the evaluation metrics. These metrics are characterized by the following equations:

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} * 100 \tag{1}$$

$$Sensitivity = \frac{tp}{tp+fn} * 100 \tag{2}$$

$$Specificity = \frac{tn}{tn+fp} * 100 \tag{3}$$

$$MCC = \frac{(tp \times tn) - (fp \times fn)}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \tag{4}$$

Where tp denotes true positive, and tn, fp, fn denote true negative, false positive, and false negative, respectively.

Comparison with Different Machine Learning Techniques

Since DeepAmp is the first computational model proposed to predict AMPylation PTM, it is not possible to compare model performance with any other studies. However, to investigate the effectiveness of CNN to build DeepAmp, we compare it with other ML models to solve this problem. Results achieved using DeepAmp compared to other ML models including Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), Decision Tree (DT), and K-Nearest Neighbor (KNN) using same set of features are presented in Tables 1 and 2 for 5-fold and 10-fold cross-validations, respectively. We present the average of 10 runs of 5-fold and 10-fold cross-validations model for all the metrics in Tables 1 and 2. As shown in these tables, DeepAmp achieves significantly better results in terms of all four metrics than other machine learning methods which are investigated in this study.

As shown in Table 1, DeepAmp achieves 73.9%, 78.6%, 71.2%, and 0.49 in terms of Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) for AMPylation site prediction task using 5-fold cross validation, respectively. Also, according to Table 2, DeepAmp achieves 77.7%, 79.1%, 76.8%, and 0.55 in terms of Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) for AMPylation site prediction task using 10-fold cross validation, respectively. As

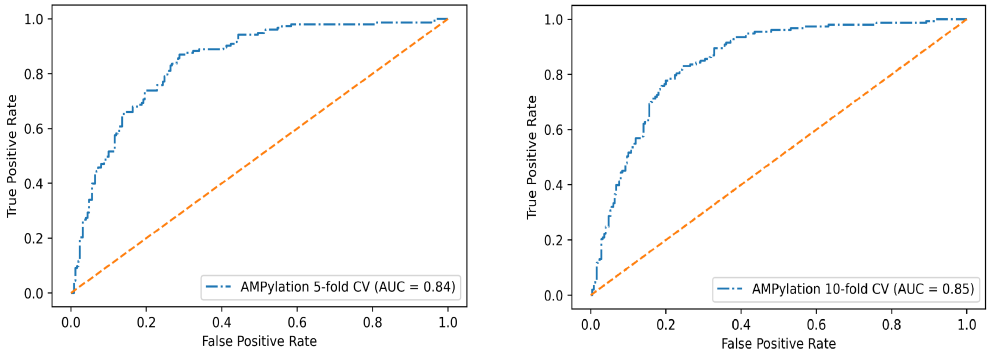


Figure 1. Receiver operating characteristic (ROC) curves for 5-fold CV and 10-fold CV DeepAmp model

Table 1. Different models result on AMPylation (5 fold CV)

Model	Accuracy	Sensitivity	Specificity	MCC
DeepAMP	**73.9**	**78.6**	**71.2**	**0.49**
LR	65.5	49.0	75.6	0.26
SVM	62.5	47.7	71.6	0.20
DT	59.6	51.0	64.8	0.16
KNN	46.4	86.3	22.0	0.10
RF	66.0	39.9	82.0	0.25

shown in these Tables, the prediction accuracy and MCC of other classifiers explored in this study are all below 70.0% and 0.30, respectively which demonstrate the effectiveness of using CNN to build DeepAmp.

As shown in Tables 1, and 2, the results using 10-fold cross-validation are slightly better than those reported using 5-fold cross-validation. This can be associated with larger number of samples used to train our model in 10-fold cross-validation. This highlight that having larger benchmark, DeepAmp is able to achieve even better results. However, the consistency between results using 5-fold and 10-fold cross-validations demonstrates the generality of DeepAmp.

In Figure 1, the receiver operating characteristic curves (ROC curves) clearly illustrate the capability of distinguishing the AMPylation and non-AMPylation sites of the DeepAmp model. Also, as shown in Tables 1, and 2, in terms of the MCC score, the other ML models display mediocre classification quality, conversely, DeepAmp shows significant improvement in the classification quality. It demonstrate the effectiveness of DeepAmp over other classifiers in identification of positive and negative samples, consistently.

Table 2. Different models result on AMPylation dataset (10 fold CV)

Model	Accuracy	Sensitivity	Specificity	MCC
DeepAMP	**77.7**	**79.0**	**76.8**	**0.55**
LR	66.2	51.0	75.6	0.28
SVM	65.0	51.6	73.2	0.26
DT	61.7	61.4	62.0	0.23
KNN	46.4	86.9	21.6	0.11
RF	67.0	40.5	83.2	0.27

Methods and Materials

This section describes the proposed method and benchmark dataset presented in this study.

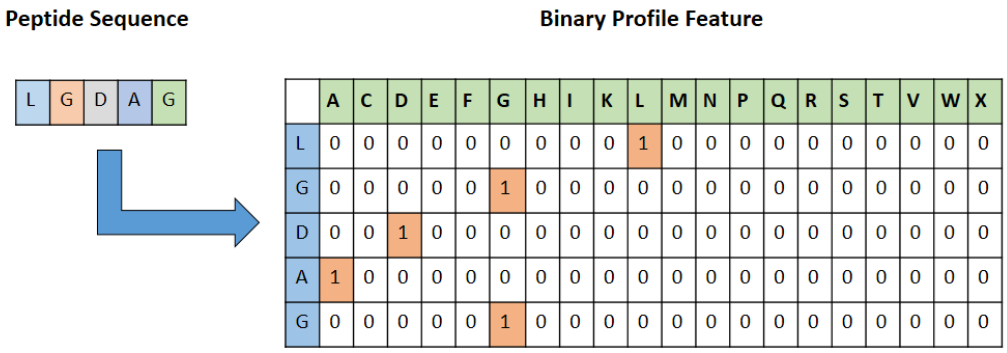
Benchmark Dataset

Kielkowski et al[9] has identified the AMPylation in intact cancer cells via LC-MS/MS as well as imaging methods. They identified a total of 162 protein sequences to be involved in this distinct modification. We investigated these proteins through UniProt database and identified a total of 133 unique protein sequences which are used to build our dataset.

We then use CD-Hit to remove proteins with over 40% sequential similarities to discard redundancy in the dataset[29]. The resulting dataset contains 130 unique proteins with less than 40% sequential similarities. After that, for each AMPylation and non AMPylation site, a 31-residue peptide containing central AMPylation /non AMPylation site with 15 residues upstream and 15 residues downstream was extracted. We tried different length of peptide-containing which among them, using 31-residue peptides attained the best results. To build the peptides sequence for AMPylation sites at the two ends of the proteins with less than 15 neighboring amino acids on each side, we use equalized by padding with "X" residue. As a result, a total of 153 peptides with AMPylated sites and 28872 peptides with non-AMPylated sites were extracted from 130 protein sequences. From the 28872 non-AMPylated sites, we selected 250 sequences randomly to balance our dataset having almost 2:1 ratio of negative to positive samples. Thus our final dataset of 403 peptide sequences with 153 AMPylated peptides and 250 non- AMPylated peptides was created. This dataset is available at .

Feature Encoding

Feature encoding is an important step in building an effective machine learning model. Binary profile features are straightforward, yet shown to be very effective for the prediction of different functionalities in the multi-omics dataset[30,31]. In this study, we generate Binary profiles for each peptide, by representing each amino acid as a vector of 20 dimensions in term of one hot encoding. For instance, Alanine is replaced by a 20 size one hot vector which is [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]. As a result, a sequence of length L was represented by a vector of dimensions L × 20. Considering L= 31 (length of peptides), we extract 620 features for each peptide (31 × 20). This feature encoding process is depicted in Figure 2. Considering that we use



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2. Binary Profile feature generation of peptide sequences

Convolutional Neural Network to build our model, binary profile can potentially provide extensive information to train our model.

Classification Technique

Convolutional neural network (CNN) is widely used in computational biology for predicting different biological and chemical functionalities and entities from multi-omics datasets. It has shown tremendous success in the prediction of different PTMs, cancer cell types classification tasks, origins of replication prediction, and many more[32–34]. Like any other neural network, a CNN consists of an input layer, hidden layer, and an output layer. Extracting feature maps using convolution operation makes the CNN architecture different from the regular neural nets. Unlike hidden layers of regular neural net which basically constructed by a set of fully connected neurons, the hidden layers of CNN mainly consist of a convolutional layer, pooling layer, and fully connected layer[35].

The CNN architecture we used is depicted in Figure 3. The input is the L × 20 matrix where L is the length of the protein sequence (31). We applied one-dimensional kernels to the input vectors. The output of our first 1-D convolutional layer which can also be thought of as a motif scanner is then passed to the max-pooling layer. Among the three convolutional layers we used, max-pooling was applied in the first two of them. The last convolutional layer output is directly passed to a fully connected layer and the prediction layer. Rectified Linear Unit (ReLU) was used as activation function for each intermediate layer as it is

popularly used for its simplicity and effectiveness[36,37]. In each of the convolutional layers and the fully connected layer, we used dropout to avoid overfitting[38].

Even though for computer vision problems deeper CNN models provide the best result[39], for biological sequence data which are presented in term of matrix as input, different studies have shown that increasing the depth of the convolutional layer does not necessarily lead to improvement in prediction accuracy specially for the smaller datasets similar to ours[40]. Furthermore, it reduces the chance of overfitting and requires fewer instances for training[38,41]. Considering these aforementioned issues, in this study a shallow CNN architecture is used for constructing DeepAmp.

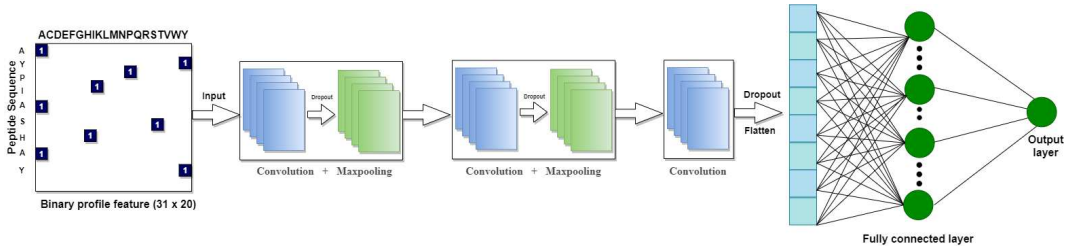


Figure 3. Model architecture of DeepAmp

Evaluation methods

In order to measure the efficacy of DeepAmp, k-fold cross-validation is used here. In k-fold cross-validation, the dataset is split into k subsets. From this k subset, k-1 is used for training and the remaining fold is used for validation. This way the whole dataset gets used for training. Since the training size gets bigger, the classifiers tend to show better results. We used stratified k-fold cross validation which maintains a fixed ratio of negative and positive sites in the training and validation dataset[42]. In this study, we evaluate our model using k = 5 and 10 as two common values for this parameter.

Conclusion

In this study, we presented a new dataset that can be used to evaluate computational methods specially machine learning based models to predict AMPylation PTM. On top of that, we proposed a new deep learning-based tool called DeepAmp for predicting AMPylation using CNN and binary profile feature vector. DeepAmp achieves an accuracy of 77.7% and sensitivity, specificity, and MCC score of 79.1%, 76.8%, 0.55, respectively for 10-fold cross-validation. DeepAmp also significantly outperforms widely used machine learning models including Support Vector Machine, K-nearest Neighbor, and Random Forest for predicting AMPylation sites. Due to the limitation of the sample size available, prediction with high accuracy is strenuous. In the future refinement of our work, we aim to incorporate new AMPylation sites into the dataset and create a larger database for AMPylation PTM. Furthermore, we aim to ameliorate our predictor's performance by using different feature sets and deeper CNN architectures. Our presented dataset and DeepAmp as an standalone predictor are publicly available at https://github.com/MehediAzim/DeepAmp.

References

1. Jensen, O. N. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. opinion chemical biology* **8**, 33–41 (2004).
2. Kia-Ki, H. & Martinage, A. Post-translational chemical modification (s) of proteins. *Int. journal biochemistry* **24**, 19–28 (1992).
3. Lin, J., Liang, H., Yan, J. & Luo, L. The molecular mechanism and post-transcriptional regulation characteristic of tetragenococcus halophilus acclimation to osmotic stress revealed by quantitative proteomics. *J. proteomics* **168**, 1–14 (2017).
4. Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of g-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol. & therapeutics* **103**, 21–80 (2004).
5. Chauhan, M., Tarique, M., Sourabh, S. & Tuteja, R. Overview of posttranslational modifications of biochemically characterized plasmodium falciparum helicases. In *Helicases from All Domains of Life*, 113–124 (Elsevier, 2019).
6. Yarbrough, M. L. *et al.* Ampylation of rho gtpases by vibrio vops disrupts effector binding and downstream signaling. *science* **323**, 269–272 (2009).

7. Yarbrough, M. L. & Orth, K. Ampylation is a new post-translational modification. *Nat. chemical biology* **5**, 378–379 (2009).
8. Casey, A. K. & Orth, K. Enzymes involved in ampylation and deampylation. *Chem. reviews* **118**, 1199–1215 (2018).
9. Kielkowski, P. *et al.* Ficd activity and ampylation remodelling modulate human neurogenesis. *Nat. communications* **11**, 1–13 (2020).
10. Mullard, A. Examining the fic domain. *Nat. Rev. Microbiol.* **7**, 405–405 (2009).
11. Itzen, A., Blankenfeldt, W. & Goody, R. S. Adenylylation: renaissance of a forgotten post-translational modification. *Trends biochemical sciences* **36**, 221–228 (2011).
12. Anderson, W. B. & Stadtman, E. Glutamine synthetase deadenylylation: A phosphorolytic reaction yielding adp as nucleotide product. *Biochem. biophysical research communications* **41**, 704–709 (1970).
13. Rahman, M. *et al.* Visual neurotransmission in drosophila requires expression of fic in glial capitate projections. *Nat. neuroscience* **15**, 871 (2012).
14. Ham, H. *et al.* Unfolded protein response-regulated drosophila fic (dfic) protein reversibly ampylates bip chaperone during endoplasmic reticulum homeostasis. *J. Biol. Chem.* **289**, 36059–36069 (2014).
15. Sieber, S. A., Cappello, S. & Kielkowski, P. From young to old: Ampylation hits the brain. *Cell Chem. Biol.* (2020).
16. Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat. reviews Mol. cell biology* **15**, 536–550 (2014).
17. Nishida, Y. *et al.* Sirt5 regulates both cytosolic and mitochondrial protein malonylation with glycolysis as a major target. *Mol. cell* **59**, 321–332 (2015).
18. Du, Y. *et al.* Lysine malonylation is elevated in type 2 diabetic mouse models and enriched in metabolic associated proteins. *Mol. & Cell. Proteomics* **14**, 227–236 (2015).
19. Xie, Z. *et al.* Lysine succinylation and lysine malonylation in histones. *Mol. & Cell. Proteomics* **11**, 100–107 (2012).
20. Olsen, C. A. Expansion of the lysine acylation landscape. *Angewandte Chemie Int. Ed.* **51**, 3755–3756 (2012).
21. Azim, S. M., Haque, M. R. & Shatabda, S. Oric-ens: A sequence-based ensemble classifier for predicting origin of replication in s. cerevisiae. *Comput. Biol. Chem.* **92**, 107502 (2021).
22. Peng, C. *et al.* The first identification of lysine malonylation substrates and its regulatory enzyme. *Mol. & cellular proteomics* **10**, M111–012658 (2011).
23. Hirschey, M. D. & Zhao, Y. Metabolic regulation by lysine malonylation, succinylation, and glutarylation. *Mol. & Cell. Proteomics* **14**, 2308–2315 (2015).
24. Reddy, H. M. *et al.* Glystruct: glycation prediction using structural properties of amino acid residues. *BMC bioinformatics* **19**, 55–64 (2019).
25. Dipta, S. R. *et al.* Semal: Accurate protein malonylation site predictor using structural and evolutionary information. *Comput. Biol. Medicine* **125**, 104022 (2020).
26. Chandra, A. *et al.* Phoglystruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids. *Sci. reports* **8**, 1–11 (2018).
27. Uddin, M. R. *et al.* Evostruct-sub: An accurate gram-positive protein subcellular localization predictor using evolutionary and structural features. *J. theoretical biology* **443**, 138–146 (2018).
28. Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y. & Campbell, M. P. Sprint-gly: Predicting n-and o-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics* **35**, 4140–4146 (2019).
29. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
30. Yi, H.-C. *et al.* Acp-dl: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther. Acids* **17**, 1–9 (2019).
31. Xiao, X., Shao, S., Ding, Y., Huang, Z. & Chou, K.-C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino acids* **30**, 49–54 (2006).
32. Le, N. Q. K., Ho, Q.-T., Nguyen, T.-T.-D. & Ou, Y.-Y. A transformer architecture based on bert and 2d convolutional neural network to identify dna enhancers from sequence information. *Briefings Bioinforma.* (2021).

33. Bhinder, B., Gilvary, C., Madhukar, N. S. & Elemento, O. Artificial intelligence in cancer research and precision medicine. *Cancer Discov.* **11**, 900–915 (2021).

34. Oberti, M. & Vaisman, I. I. cnnalpha: Protein disordered regions prediction by reduced amino acid alphabets and convolutional neural networks. *Proteins: Struct. Funct. Bioinforma.* **88**, 1472–1481 (2020).

35. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. neural information processing systems* **25**, 1097–1105 (2012).

36. Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks* **94**, 103–114 (2017).

37. Li, Y. & Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. *arXiv preprint arXiv:1705.09886* (2017).

38. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal machine learning research* **15**, 1929–1958 (2014).

39. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).

40. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Briefings bioinformatics* **18**, 851–869 (2017).

41. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456 (PMLR, 2015).

42. He, H. & Ma, Y. Imbalanced learning: foundations, algorithms, and applications. (2013).

Author contributions statement

S. M. Azim designed and performed the experiments. S. M. Azim, A. Sharma, S. Shatabda, and A. Dehzangi wrote the manuscript. S. M. Azim and A. Dehzangi helped with figures and literature review. A. Sharma, A. Dehzangi, and S. Shatabda mentored and analytically reviewed the paper. All the authors reviewed the article.